AD_____

Award Number: W81XWH-09-1-0546

TITLE: Oncogenicity and Selective Inhibition of ERG Splicing Variants in Prostate Cancer

PRINCIPAL INVESTIGATOR: Luca Cartegni, Ph.D.

CONTRACTING ORGANIZATION: Memorial Sloan Kettering Cancer Center
New York, NY 10025

REPORT DATE: March 2012

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| March 2012 | Final | 1 September 2009 – 29 February 2012 |

**4. TITLE AND SUBTITLE**

Oncogenicity and Selective Inhibition of ERG Splicing Variants in Prostate Cancer

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
W81XWH-09-1-0546

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Luca Cartegni, George Boutsalis, Gina Rocco, Francesca Zammarchi

**E-Mail:** Cartegnl@mskcc.org

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Memorial Sloan Kettering Cancer Center
New York, NY 10025

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

ERG, a member of the ETS transcription factor family, is frequently overexpressed in prostate cancer as a result of its fusion to the androgen-responsive Tmprss2 gene. Different genomic rearrangements and alternative splicing events around the junction region lead to multiple combination of Tmprss2:ERG fusion transcripts that correlate with different tumor aggressiveness, but their specific functions and biological activities are still unclear. The complexity of ERG expression pattern is compounded by the use of alternative promoters, splice sites, polyadenylation sites and translation initiation sites in both the native and fusion contexts. Our systematic characterization of native ERG and Tmprss2:ERG variants reveals that their different oncogenic potential is impacted by the status of the Ets domain and the configuration of the 5' UTR region. In particular, expression and activity of functional ERG and Tmprss2:ERG variants are influenced both by translation initiation signals within the different isoforms and by inhibitory upstream Open Reading Frames (uORF) in their 5' UTRs. Stable expression of ERG and Tmprss2:ERG variants promoted cell migration/invasion, induced a block of proliferation and induced a senescence-like state, suggesting a role for these variants in the prostate tumorigenesis process. In addition to Tmprss2:ERG fusion products, a group of related native ERG isoforms is also highly over-expressed in fusion-carrying prostate cancers, and share the same translation initiation site (in ERG exon 4) with the commonly observed Tmprss2 exon1 joined to ERG exon 4 (T1:E4) fusion-derived variant. Usage of this ATG can be preferentially down-regulated by directed antisense-based compounds, possibly representing the basis of a targeted approach that distinguishes between tumor–associated and normal ERG.

**15. SUBJECT TERMS**
tmprss2, ERG, prostate cancer, alternative splicing, translational control, antisense

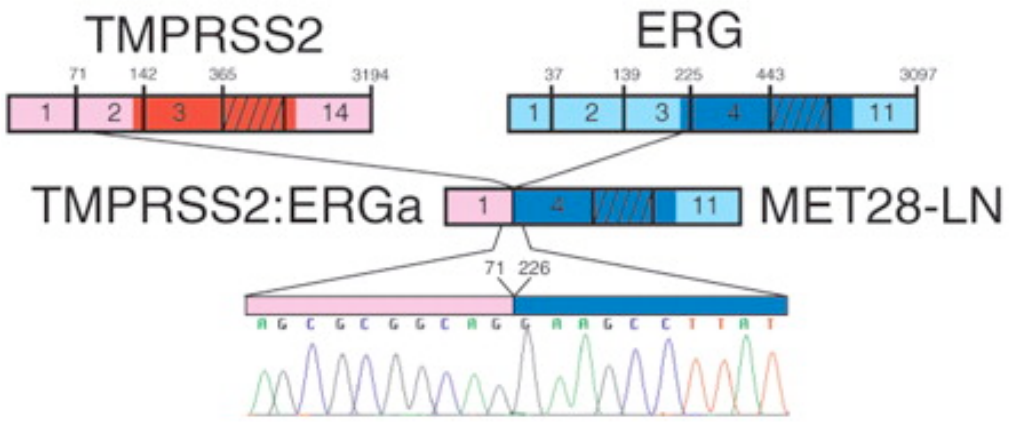
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | USAMRMC |
| U | U | U | UU | 34 | 19b. TELEPHONE NUMBER *(include area code)* |

## Table of Contents

## INTRODUCTION

A well established key event in the development haematological malignancies and sarcomas is the occurrence of chromosomal translocations [1]. Until recently, such rearrangements were not considered to play a major role in carcinomas of epithelial origin. However, the prostate-specific identification of frequent recurrent translocations between the androgen-responsive TMPRSS2 gene and members of the Ets family of transcription factors, has forced to revisit this notion and has changed the panorama of prostate cancer biology [2]. As Ets proteins are already implicated in other oncogenic translocations, the likely consequence of their fusion to androgen responsive genes would be the acquisition of the tumorigenic properties associated to the Ets transcription factors by cells sensitive to androgen stimulation, such as prostate cells.

Prostate cancer is the most common malignancy in men in developed countries, and the leading cause of cancer-related death in males [3]. More than 80% of prostate cancers harbor fusions which typically involve the 5' region of the androgen-responsive TMPRSS2 locus (including its promoter) joined to the 3' region of various Ets genes (lacking the promoters but including all or most of the Open Reading Frame, ORF) [4]. ETV1 and ERG were the first identified 3' fusion partners of the TMPRSS2 gene [2,5], but subsequent analysis lead to the description of additional Ets family members as 3' partners for TMPRSS2 and of other 5' partners for Ets genes[5,6,7]. The single most common rearrangement, TMPRSS2-exon1:ERG-exon4 (T1:E4, or variant III) arises in ~50% of prostate cancer cases[4] as a fusion that joins the TMPRSS2 5' UTR to most of ERG ORF (Figure 1).




***1. TMPRSS2/ERG fusion.*** *A common translocation joins introns 1 and 3 of TMPRSS2 (pink) and ERG (blue), respectively, generating the depicted fused transcript in over 50% of tested prostate cancer samples. The darker boxes represent coding regions.*

Multiple observations suggest that Ets gene fusions may play a role in the transition from prostatic intraepithelial neoplasia (PIN) to adenocarcinoma and invasion, and are associated to aggressive lesions and poor prognosis [8,9,10,11,12]. Overexpression of ERG in prostate cell lines activates cell invasion programs and results in the development of PINs in mice, but it is not sufficient to drive carcinogenesis [6,9,11,13]. Cooperation with separate genetic lesions, such as for example pTEN loss, that dysregulate cellular proliferationa and other control mechanisms is needed to trigger progression to advanced disease [8,14,15].

Several variants of the normal ERG gene product have been described, arising from a combination of alternative splicing, polyadenilation and transcriptional initiation[16,17,18,19]. The encoded ERG protein isoforms interact with the AP1 complex to activate transcription and their activity is modulated by homo- and heterodimeric interactions among ERG and other Ets variants [20,21]. Variability in the coding region can influence ERG activity of the TMPRSS2:ERG fusions as well, and the presence of a variant including a 72 nt alternative exon shows enhanced biological activity, especially when expressed together with other isoforms [21].

The specifics of the genomic rearrangements also introduce considerable structural heterogeneity in the 5' region. In addition to the common T1:E4 transcript, Tmprss2 exons 1, 2 and 3 can be combined with ERG exons 2, 3, 4, 5 and 6 in various alternative splicing patterns that can generate at least 17 distinct Tmprss2:ERG transcripts[2,22,23,24]. These can serve as markers for disease progression and correlate to the aggressiveness of the tumors[9,10,22]. In particular, the T2:E4 variant (Tmprss2-exon2:ERG-exon4 or variant VI), where the native ATG in Tmprss2 exon 2 is in frame with the ERG ORF, is associated with pathological and clinical aspects of aggressive disease [22].

The mechanistic basis for the different oncogenic potential of the fusion isoforms remains to be elucidated, and it could be related to intrinsic differences in the N-terminal regions or to the effect(s) that variation in the 5' region of the mRNA can have on RNA stability or expression.

To evaluate and characterize the various ERG and TMPRSS2:ERG isoforms, we sought to assess differences between the use of 3 alternative promoters, 2 common alternative splicing events and 3 polyadenylation sites (PAS) in normal tissues relative to TMPRSS2:ERG-expressing prostate tumors. These independently regulated events combine to generate 30

‘main’ native isoforms, some of which are also highly overexpressed in tumors. The characterization of the translation initiation sites used by the most common native ERG and fusion variants reveals the specific organization of the 5’ UTR region as one of the principal determinants of their biological activity and identifies an ATG in exon 4 as a promising target for antisense-based translation inhibition in prostate cancer.

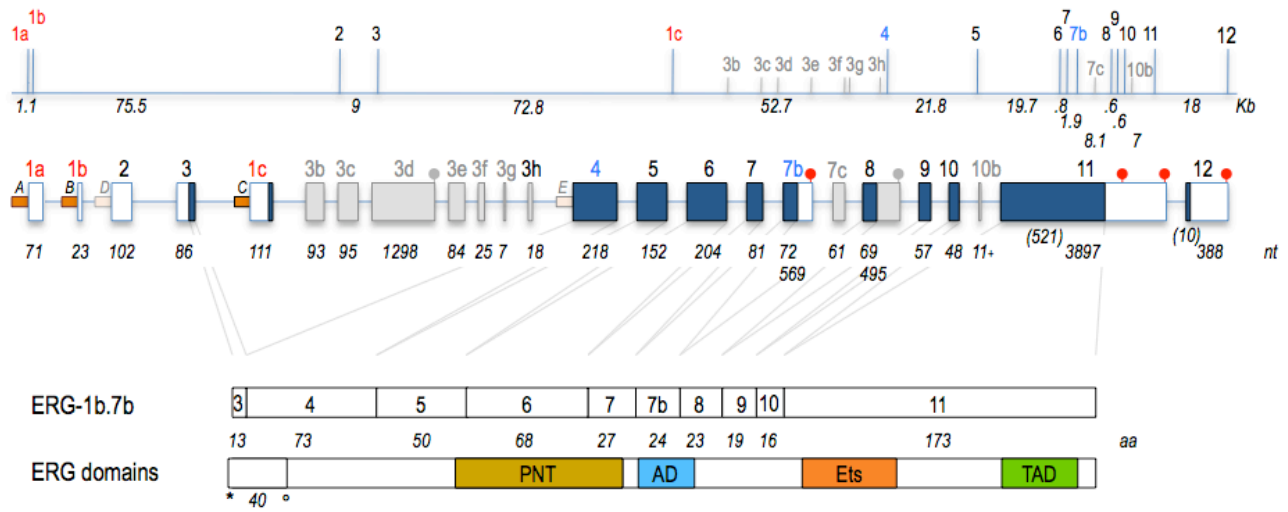
The development of specific translation-blocking compounds that can effectively and selectively reduce the levels of aberrant ERG isoforms would introduce an important set of tools to enhance our understanding of a pathway that is improperly activated in the majority of prostate cancer occurrences, and could form the basis for novel approach in their treatment.

### *Structure of the ERG gene*

Multiple ERG isoforms can arise from the human ERG gene due to a combination of alternative transcription initiation sites, splicing and polyadenylation. These isoforms can in turn combine with TMPRSS2 and other 5' partners to produce a large number of ERG-derived variant mRNAs, with variable prognostic values [9,10,22]. In order to better understand the activities of various ERG-derived oncogenic products, we sought to initially clarify the ERG gene structure and to characterize the expression patterns of the variants expressed.

This is very complex, and one problem consists in the considerable contradictions in the ERG nomenclature, particularly regarding the identity of specific isoforms and their exons, with at least four different classification schemes. For example the large terminal exon that contains the Ets and transactivation domains is variably referred to as exon 11[2,25], 12 [4], 16 [19,26], 17 [21]. Similar discrepancies are also found when mRNAs or protein isoforms are involved, with the obvious consequence of generating confusion when trying to interpret the results of different groups which might adopt different conventions.

We worked out and describe in Figure 2 an up-to-date view of the exon-intron structure of the ~300 KB ERG locus (ENSG000000157554) and propose a unified, rational nomenclature




***2. Structure of ERG gene.*** *The top structure indicates the genomic organization of the gene, encompassing ~300 KB. Introns are drawn roughly to scale and exon are indicated by vertical bars. Intron sizes of the main variants are indicated. Red lettering indicates First exons, and blue lettering the main alternative splicing events. The next scheme indicates a more detailed representation of constitutive and alternative exons, with white boxes indicating untranslated regions and blue boxes indicating coding regions. Grey exon are novel putative, low abundance exons. The orange boxes indicate promoters and the red circles mark the principal polyA sites. On the bottom, the protein domain structure is schematized with the corresponding encoding exons. The star and empty dot on the left indicates alternative initiation ATGs.*

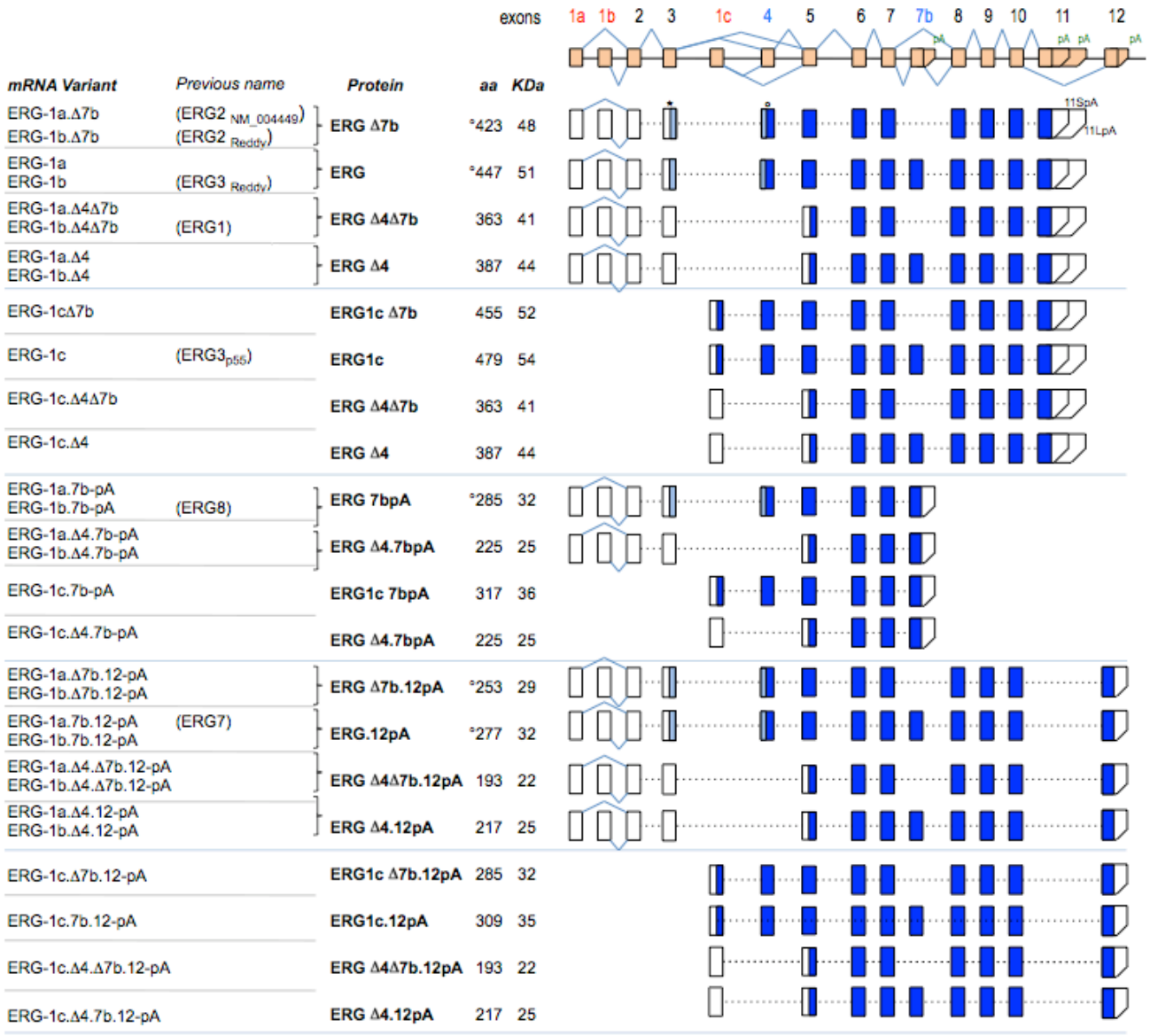
that aims at incorporating the most established conventions based on prevalent literature, in order to minimize confusion.

The structure is mainly based on the 11 exons of the ERG Refseq entry NM_004449.4 (corresponding to Uniprot entry P11308 for ERG2), and incorporates the exon numbering used in the seminal Tomlins paper first describing the TMPRSS2:ERG fusion [2,25]. In short, we maintained as “exon 4” the 218nt exon that is the main partner of Tmprss2 and as “exon 11” the large 3897nt exon encoding the Ets domain; we named Exon 1a, 1b, 1c the three mutually exclusive ‘first’ exons following the three validated promoters $P_A$, $P_B$ and $P_C$. (of note, Exon 1c is contained within intron 3 and splices directly to exon 4); To ensure the rational inclusion of additional exons that might be identified in the future, we distinguished by letters the alternative exons not part of the 11 reference ones. For example, the 72nt alternative exon included between exons 7 and 8 is called ‘exon 7b’, and so on.

We then identified the main alternative regulatory events that generate the majority of ERG variability: 3 alternative promoters (PA, PB and PC); two common alternative splicing events (inclusion/skipping of exon 4 and 7b); three separate polyadenylation sites (PAS 7bpA, 11LpA and 12pA). Combinatorial usage of the compatible alternative events generates 30 possible ‘major’ ERG transcript variants (Fig 3), which can encode 15 different predicted ERG-related polypeptides, with 3 different N-terminals, 3 different C-terminals and 1 possible internal variation (inclusion or skipping of 24 aa in the Alternative Domain encoded by Exon 7b).

In addition to the 30 ‘major’ variants described in figure 1, a plethora of ‘minor’ isoforms heve been reported in literature or are present in databases (Suppl Fig 1). The list includes variants showing skipping of exons 2, 5, 7 and 8; usage of a proximal (Short) polyA site in exon 11 (11SpA) or of an additional intronic one downstream of exon 8 (8pA); and inclusion of supplementary alternative exons 7c, 10b and of multiple alternative exons in intron 3 (exons 3b-h), indicated in gray in figure 1. The evidence for the specific size of some of these exons is non conclusive as they seem to derive from partial cDNAs that start or end within the exons themselves. Isoforms ERG4, ERG6 and ERG9 from the previous Owczarek study [19] are minor variants and have been reclassified in this group, whereas ERG 5 appears to be a truncated cDNA derived from one of the main isoforms.

Because any of these additional minor events could in principle combine independently with all the structurally compatible major isoforms, hundreds of variants could potentially be




***3. Main ERG variants.*** ERG three promoters (pA, pB, pC), two main alternative splicing events (skipping/inclusion of exons 4 and 7b) and three main polyA sites (7bpA, 11LpA, 12pA) combine to generate 30 principal mRNA variants, which can encode 15 different ERG protein isoforms. The new proposed nomenclature, together with the previous known of the few descrived variants, the protein encoded, its size and predicted MW are indicated, along with a cartoon of the exon structure. Usage of exon 1a or 1b is always mutually exclusive and depends on the promoter engaged, and it always splices to exon 2, as indicated by the light blue chevrons.

generated. However, all these events appear to occur very sporadically, and their physiological relevance is so far unknown.

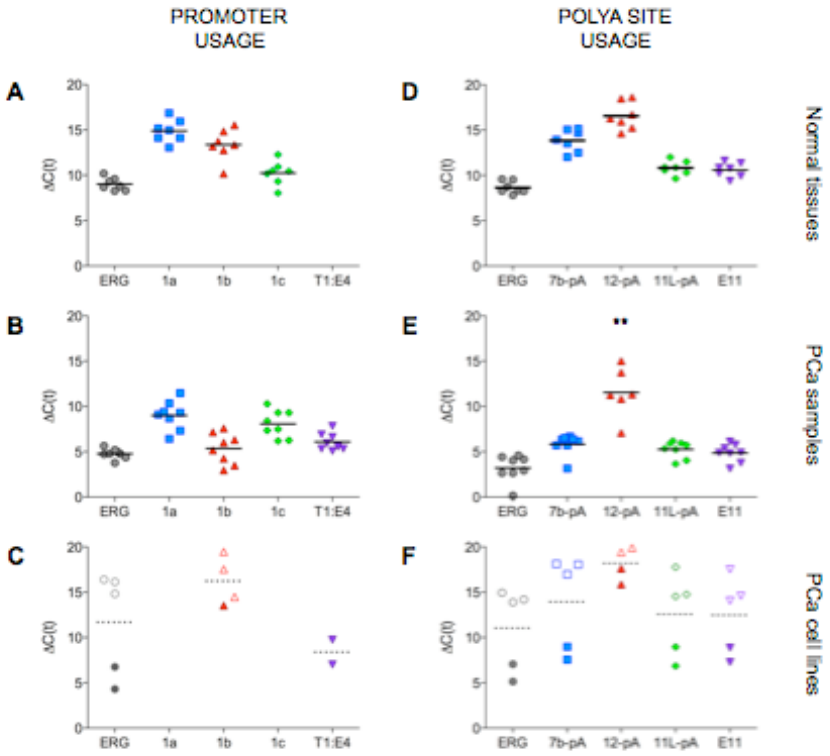
The structure of the ERG gene was much more complex than originally anticipated, and its full delucidation reported here was a major undertaking that took significant amounts of time and resources.

We have then proceeded to carry on characterization of only what appear to be the main variants, which are more likely to be of physiological significance in normal ERG functions and in disease

### *ERG expression: a cancer-associated switch in promoter usage*

We set out to investigate usage of promoters, PAS or splicing signals by qPCR analysis of ERG expression in different normal tissues, tumors and cell lines.

While some degree of tissue to tissue variability is observed (Suppl Fig 2), in general in normal tissues promoter PC (mean Dc(t)= 10.2) appears to be the most active, being ~25-fold and ~10-fold more active than promoters PA (Dc(t)= 14.9) and PB (Dc(t)= 13.4), respectively (Figure 4A, lower  point means higher expression levels). On the other hand, in a panel of 8 primary PCa samples expressing TMPRSS2-ERG fusions, promoter PB (mean Dc(t)=5.4) accounts for the majority of native ERG transcript and it is present at levels comparable to those




***4. Expression of ERG isoforms***. *(**A-C**) Expression levels of total ERG, and specific isofroms generated from promoters pA, pB and pC were measured by qPCR with specific primers, along with the expression of the fusion T1:T4 variant,in normal tissues (A), prostate tumor samples (B) and prostate cancer cell lines (C). (**D-F**) Expression levels of total ERG, and specific isofroms generated from polyA sites 7bpA, 12pA, 11LpA were measured by qPCR with specific primers, along with the expression of the exon 11, in normal tissues (A), prostate tumor samples (B) and prostate cancer cell lines (C). The indicated values in the graphs represent averages of at least 3 independent experiments and are presented as ΔC(t) normalized to GAPDH housekeeping gene, therefore a "high" ΔC(t) value means low levels of expression and a "low" value means high level of expression. See supplemental figure 2 and 3 for additional details.*

of the fusion itself (Dc(t)=6), over 100-fold more abundant than the same transcript in normal prostate (Dc(t)= 12.8) (Figure 4B and SupplFig 2A).

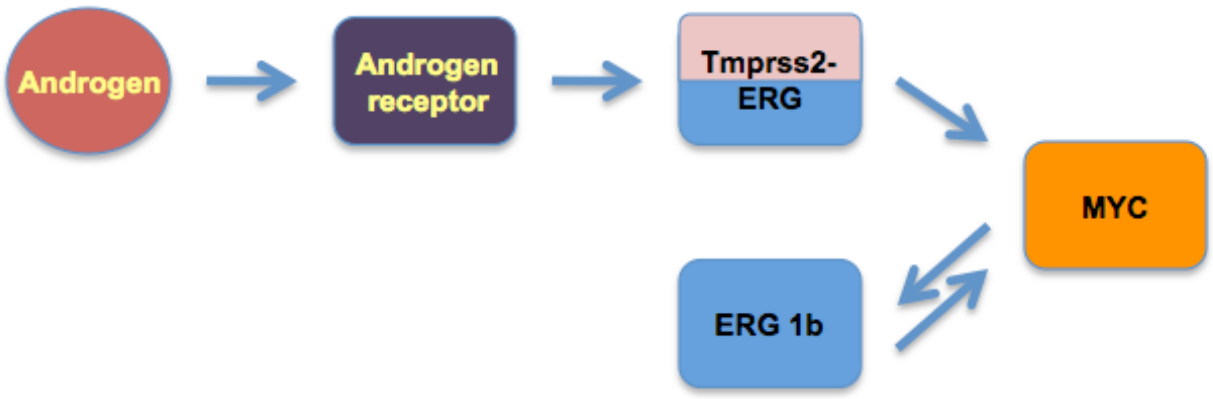
Indeed, the native PB promoter appears to be the principal source of ERG transcript in at least some of the TMPRSS2:ERG carrying samples (SupplFig 2). Promoters PA (Dc(t)=9) and to a lesser extent PC (Dc(t)=8) are also activated in tumors when compared to normal prostate, but not as much as PB. In PCa cell lines, the preferential activation of PB is even more pronounced, as it is the only native ERG promoter active, while signals from PA and PC are undetectable (Figure 4C).

Two of the prostate cancer cell lines tested, VCap and NCI-H660 carry the TMPRSS2:ERG fusion. Accordingly, the fusion transcripts are very abundant (full symbols) in their RNA, while they are absent in that from the other PCa cell lines that do not carry the fusion (LNCap, DU145, C4-2, open symbols and SupplFig 2). As expected, no expression from any of the native promoters was observed in the NCI-H660 cell line, which carries the fusion on both alleles, with loss of all the endogenous promoters [27].

From this first set of experiments we conclude that a promoter switch occurs in PCa tumors and cell lines and that increased usage of native promoter $P_B$ is associate with (and possibly contributes to) the tumor phenotype.

Overexpression of ERG in PCa was previously described [26], but following the discovery of the Tmprss2:ERG fusion, it has been typically ascribed to this event. However, our finding that in addition to the fusion-derived transcripts, some native ERG variants can also be highly overexpressed in PCa, suggests a bigger role for native ERG in PCa development. This is supported by the observation that endogenous mouse *Erg* transcripts are overexpressed in tumors from prostate conditional *Pten*$^{-/-}$;*Trp53*$^{-/-}$ mice, compared to *Pten*$^{-/-}$;*Trp53*$^{+/+}$ mice [8]. Importantly, while the latter model results in an indolent form of PCa, the former produces an aggressive phenotype [8].

The differential activation of the three native promoters in the fusion-carrying PCa samples suggests that this aberrant expression is transcriptionally regulated. An intriguing possibility is that activation of the native $P_B$ promoter may be driven directly or indirectly by ERG itself as part of a positive regulatory loop. For example, several putative c-Myc responsive elements were identified immediately upstream of the ERG $P_B$ promoter [19]. Since c-Myc is a key downstream target of ERG [28], the androgen-dependent activation of ERG from the fusion, or a separate




***5. Mechanism of MYC-dependent activation of endogenous ERG genes .*** *Androgen-activated androgen receptor stimulates Tmprss2 promoter, thus inducing expression of the tmprss2-ERG fusion. Erg stimulates expression of MYC, which in turns promotes specific upregulation of the ERG 1b variant, triggering a positive-feedback loop that sustains MYC expression and could in principle become androgen independent*

PTEN-dependent Myc activation [29] could trigger a self–sustaining ERG/Myc oncogenic loop, which could eventually become androgen-independent. Indeed, consistently with this model, a feed-forward mechanism where expression of endogenous ERG is controlled by overexpression of the fusion product has recently been described [30].

### *ERG expression: alternative polyadenylation and splicing*

Of the three principal polyadenylation signals described (Figure 2), the exon 11 Long polyA site (11L-pA) is needed to generate a fully functional ERG protein. The other polyadenylation sites in intron 7b (7b-pA) and exon 12 (12-pA) give rise to C-terminally truncated ERG isoforms lacking the functional Ets domain either because the transcripts terminate early (7b-pA) or because exon 11 is skipped when exon 12 is used (12-pA). In normal tissues, the 11L-pA signal generating the full-length transcript, was the most commonly used (mean $\Delta C(t)=10.8$), about 8-fold more than 7b-pA ($\Delta C(t)=13.8$) and 50-fold more than 12-pA $\Delta C(t)=16.6$) (Fig 4D).

Interestingly, in the TMPRSS2:ERG-expressing prostate tumors, the usage of 7b-pA, results strongly activated and is about as common as 11L-pA (Figure 4E). The same is true in PCa celll lines expressing the fusion (Figure 4F) suggesting that transcripts from the TMPRSS2:ERG fusion preferentially use the 7b-pA site, and that switching to this pattern of expression could be associated to tumor progression. A second, proximal polyA site producing a shorter 3' UTR had

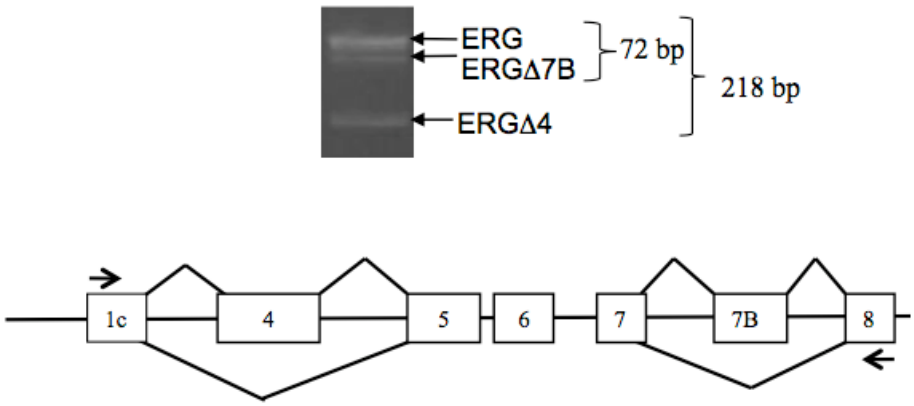



***Figure 6 Alternative Splicing hot spots in ERG gene.*** *PCR reaction spanning exons 1c to 8, or exons 3 to 8 reveal the consistent presence of multiple bands. Sequencing of these commonly observed alternatively spliced PCR products in multiple cell lines and tissues has led to the identification of inclusion/skipping of exon 4 (218 bp) and of exon 7b (72 bp) as splicing "hot spots" in the ERG gene, in both the normal and the fusion contexts. An example of a representative PCR reaction from RNAs derived from VCap cells using primers amplifying from exons 1c and 8 is shown (depicted by arrows).*

been described for exon 11 (11S-pA) [16,19]. We indirectly assessed its use in our samples by quantitatively amplifying regions in exon 11 located before and after the proximal site (E11 and 11L-pA). Because the levels of expression in these two regions are the same in all samples analyzed, we conclude that usage of 11S-pA is marginal under most conditions (Fig 4D-F and Suppl Fig 3).

To determine if any of the alternative splicing events described above was differentially regulated in PCa, we analyzed regions around exon 7b and exon 4 in normal tissues, PCa samples and cell lines by semi quantitative-PCR. Both alternative events were readily detectable (Figure 5), with a clear prevalence of the exon 4 and exon7b inclusion variants, but we didn't observe any significant difference in the relative amount of splicing variants between normal tissue and prostate samples, suggesting that the regulation of these two specific splicing events does not play a significant role in prostate cancer.

### *N-terminal heterogeneity of ERG isoforms: mapping of ERG and Tmprss2 –ERG starting ATG.*

To evaluate whether the heterogeneity in the 5' region affected ERG expression and activity, we subcloned into a mammalian expression vector the native variants ERG-1a, ERG-1b, ERG-1c, ERG-1b.Δ4, ERG-1c.Δ4, and the common fusion variant T1:E4, with their complete 5' UTRs (Fig 7A). In all cases exon 7b was included in the cDNA, as this is the most frequent configuration.

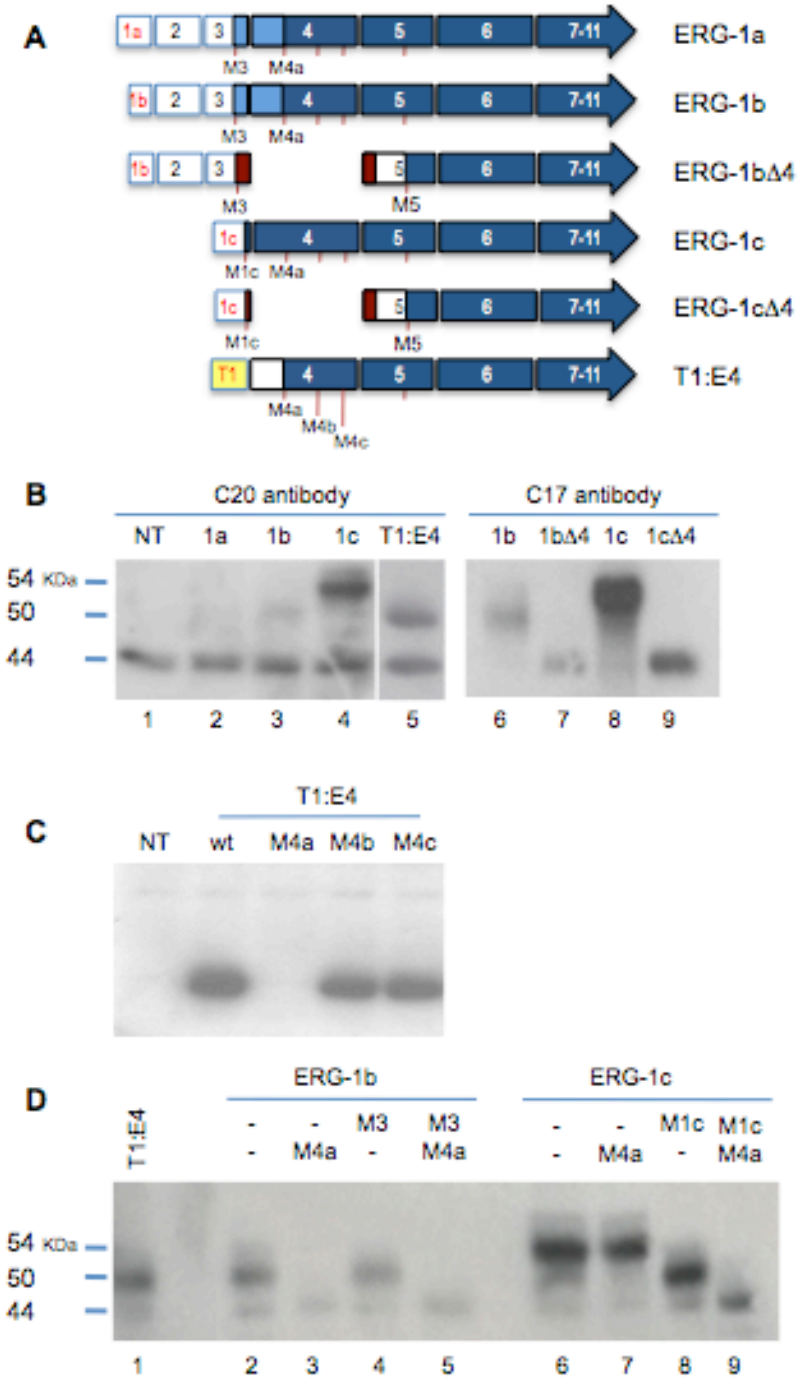



***Figure 7.*** **Mapping of translation initiation of ERG isoforms. (A)** N-terminal heterogeneity in human native ERG variants. The 5' regions for the indicated variants are reproduced. Boxes represent exons with the ORF in blue. Red ticks below the exons indicate in-frame ATGs in exons 1c and exons 3 to 5. The out-of-frame products caused by exon skipping, from the otherwise in frame ATGs are indicated in red. **(B)** Western Blot (WB) of transient expression of the variants in HeLa cells. Antibody C20 recognizes exogenous protein and an endogenous band at ~44 KDa corresponding in size to D4 variants. Antibody C17 preferentially recognizes exogenous ERG bands. **(C)** Mutations that independently eliminate the 3 in-frame ATGs in exon 4 were introduced into the cDNA of T1:E4 and analyzed as above. **(D)**, Similarly, mutations were introduced to eliminate the ATGs in exon 3 (ERG-1b) or in exon 1c (ERG-1c) by themselves or in combination with mutations in the first in-frame ATG (M4a) in exon 4 (ERG-1b and ERG-1c).

Upon transient transfection in HeLa cells, expression of ERG-1c was efficient and corresponded to a peptide of the expected 54 KDa size (Fig. 7A). On the contrary, expression of the native ERG-1a and ERG-1b variants was inefficient, and resulted in peptides migrating at a size smaller than the 55 KDa expected if the first in-frame predicted ATG in exon 3 was used (Fig 7B, lanes 2-3). Multiple explanations could account for the unexpected gel mobility, such as the N-terminal conformation affecting migration, differential processing of the unique N-terminal or use of a downstream ATG. Interestingly, this peptide co-migrated with that generated by the transient over-expression of the T1:E4 fusion variant, at around 50 KDa (lane 5).

Additional variation in the N-terminal region derives from exon 4 skipping which alters ERG's main ORF, so the predicted starting ATGs in exon 1c or exon 3 cannot generate an ERG-related peptide (Fig 7A). Transient overexpression of ERG-1b.Δ4, ERG-1c.Δ4 results in both cases in peptides migrating at around 44KDa, consistent with the predicted usage of an in-frame ATG in exon 5 (Figure 7B, lanes 6-7). This is most evident when using EG antibodty C-17 (Cell Signaling), which readily recognize recombinant ERG, but gives a very low signal of endogenous proteins from lysates from multiple cell lines. A different ERG antibody (anti-ERG C-20, Cell Signaling) preferentially recognizes in lysates an endogenous band corresponding in size to ERGD4 (lanes 1-5), but the precise identity of this product is unclear) so the switch to ATG M5 usage would be hard to detect.

Since T1:E4 lacks the predicted initiation codons from both TMPRSS2 and ERG transcripts, an alternative internal ATG from within ERG's open reading frame must be used to express an ERG-related product from the fusion mRNA, likely from exon 4. To identify the T1:E4 initiation codon, a series of methionine to alanine point mutations were generated in exon 4 of ERG for the three in-frame ATG codons encountered (Fig 7B). Mutation at nucleotides 79 (M4a), but not 121 (M4b) and 184 (M4c) of exon 4 abolished T1:E4 expression (Figure 7C), indicating that the fusion transcript uses the first in-frame ATG for expression of the ERG peptide.

Similar mutations in the first in-frame ATGs in exon 3, 1c and 4 where also introduced alone or in combination to map the starting ATGs from the native ERG isoforms (fig. 7D). Mutation of the ATG in exon 3 (M3) didn't have any effect on the expression of the 51 KDa product (Figure 7D, lanes 2 and 4), while mutation of the next ATG in exon 4 (M4a) abrogated it both in the wt and M3 context (Figure 7D, lanes 3 and 5), indicating that ERG-1b (and ERG-1a) do not use

their first in-frame ATG, as would be predicted by the 5' cap dependent scanning model of translation initiation site selection [31], but instead it selects the following ATG in exon 4, generating a peptide identical to that encoded by the T1:E4 fusion.

On the other hand, when the initial start codon in exon 1c is mutated (M1c), ERG-1c mobility is reduced from ~54 KDa to ~51 KDa (Figure 7D, lanes 6 and 8), like T1:E4 and ERG-1b (lane 1 and 2), consistent with a switch to the M4a ATG. Indeed, when this is mutated in the context of M1c (M1c/M4a) the 51 KDa product is also abrogated in favor of a smaller product, which appears to correspond to relatively inefficient usage of the ATG in exon 5. Altogether, these experiment show that native variants ERG-1c, ERG-1b.Δ4, ERG-1c.Δ4, as well as the common T1:E4 fusion, mainly use the first predicted in-frame ATG available, as expected from the scanning model. On the contrary, the ERG-1a and ERG-1b variants preferentially use the second in-frame ATG, and do so at a significantly lower efficiency.

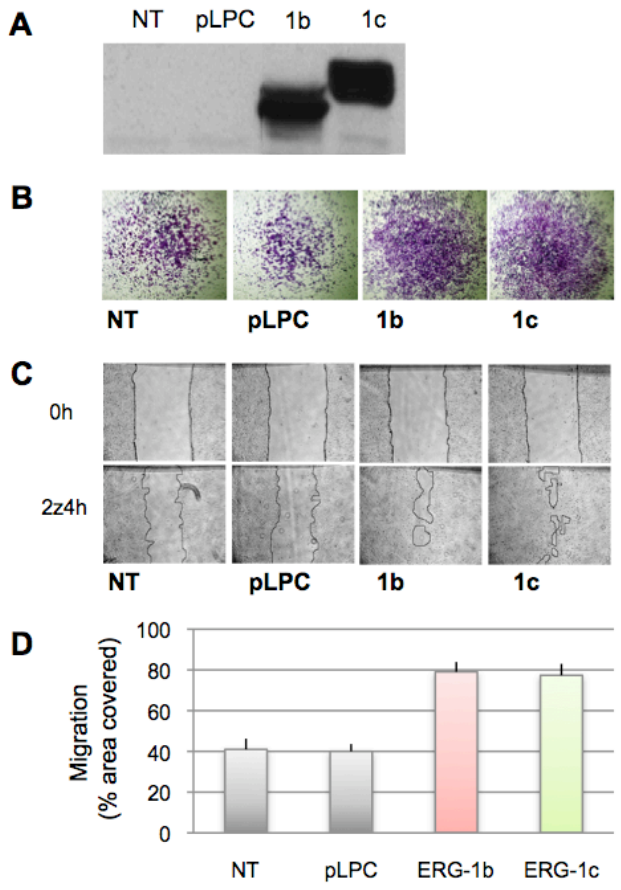



***Figure 8. Effect of expression of ERG variant in 3T3 cells.*** *Variants 1b (=Tmprss2-ERG) and 1c were stably expressed, along with controls, using retroviral vecotrs. A) western blot indicating overexpression of the isoform, Migratory properties were monitered using transwell-miration assay (B) and wound helaing assay (C). the graph below indicates a quantitation of multiple wound-healing assays*

### *ERG induces mobility but inhibits growt.h*

To assess whether the structural N-terminal differences between the cancer-associated ERG-1b/T1:E4 and the normal ERG-1c affect intrinsically their biological activity, we initially stably expressed their corresponding cDNAs in NIH-3T3 cells (Figure 8A) and selected clones with robust expression. In agreement with previous reports, we observed promotion of cell invasion and migration by both ERG variants, as assayed by trans-well migration through matrigel (Figure 8B) and scratch wound assay (Figure 8C-D). The extent of the effect on migration/invasion was comparable in ERG-1b and ERG-1c indicating

that they are similarly active, at least in some basic aspects of their biology.

Surprisingly, expression of ERG variants in this context resulted in decreased cellular proliferation (Figure 9A). This was not associated with cell death. On the contrary, expression of ERG-1b and especially ERG-1c resulted in protection from cell death following serum starvation (Figure 9B). This behavior could reflect the activation of an oncogene-dependent senescence-like state by ERG expression as previously described for other oncogenes [32]. Indeed, ERG-expressing cells, but not controls, stained positive for senescence biomarker (beta)-Galactosidase (Figure 9C). None of the isoforms, however, was sufficient to mediate cellular transformation and promote anchorage-independent growth in soft-agar (not shown).

Intrigued by these results we subsequently stably overexpressed the ERG-1b/T1:E4 isoform and the corresponding empty vector in the normal human fibroblast cell line IMR90, a well-characterized cell model system for cellular senescence. As with the NIH-3T3 cells, over-

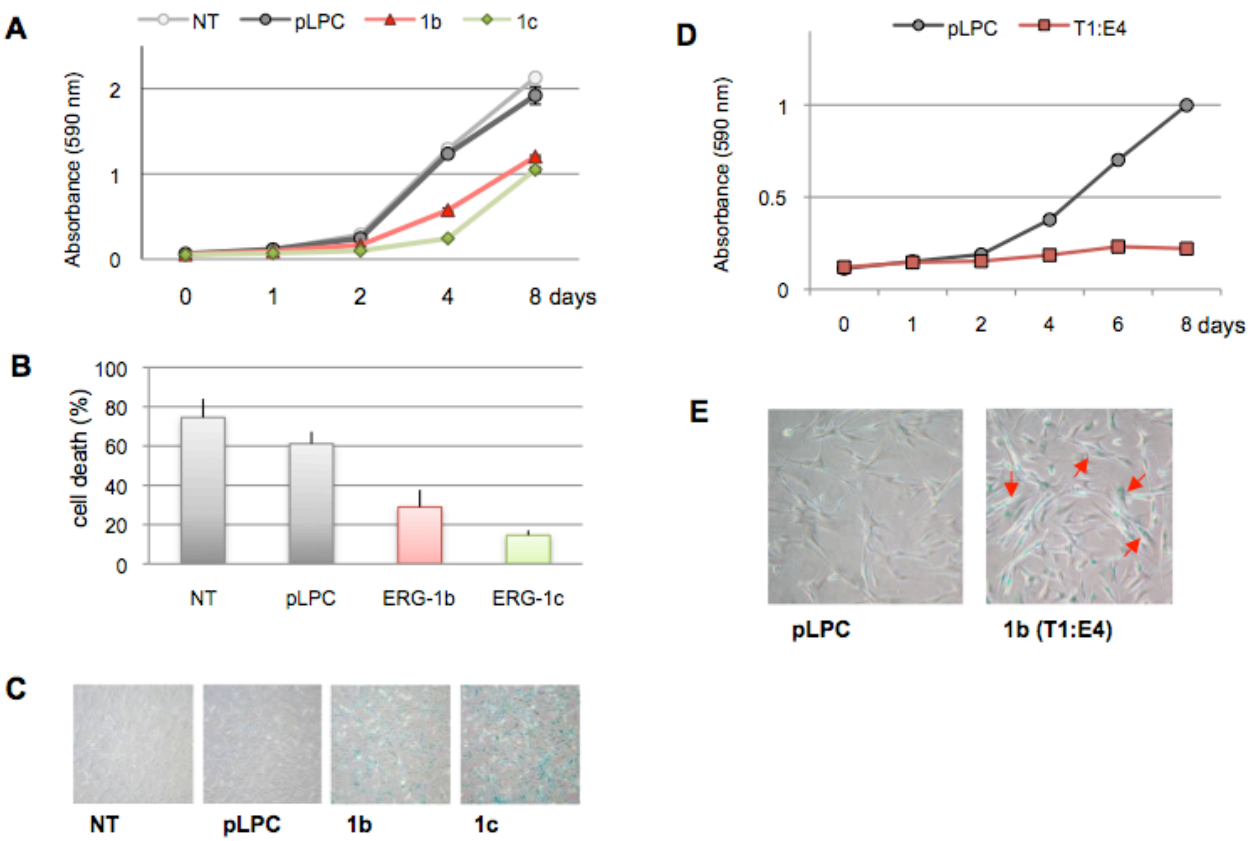



***Figure 9. Effect of expression of ERG variants.*** *Variants 1b (=Tmprss2-ERG) and 1c were stably expressed, along with controls, using retroviral vectotrs.* ***(A)*** *Growth curve of NIH-3T3 stable clones. Following selection, cells were fixed at 48 h intervals over a period of 8 days and stained with crystal violet.* ***(B)*** *After drug selection, NIH-3T3 stable clones were plated and serum starved, and cell death was measured by trypan blue method.* ***(C)*** *5 days after drug selection NIH-3T3 stable clones were stained for SA-b-galactosidase.* ***(D)*** *Growth curve of IMR90 cells stably expressing high-level of ERG1b/T1:E4 or the empty vector. After drug selection, cells were fixed at 48 h intervals over a period of 8 days and stained with crystal violet.* ***(E)*** *5 days after drug selection IMR90 stable clones were stained for SA-b-galactosidase. Red arrows indicate bi-nucleated cells.*

expression of the ERG-1b/T1:E4 isoform resulted in increased cell migration and cell invasion, (Supplemental figure 4). Moreover, IMR90 cells over-expressing the ERG-1b/T1:E4 isoform showed senescence-like phenotypes such as reduced cellular proliferation (Figure 9D), accumulation of bi-nucleated cells and elevated SA-β-gal activity, a classical biomarker of senescence (Figure 9E).

This result is particularly exciting in light of the fact that the senescence program is activated once a cell has sensed a critical level of damage or dysfunction, pointing to a critical role for ERG-1b/T1:E4 overexpression as an initial event leading to PCa. Moreover it is important to note that cellular senescence can be detected in early-stage human PCa specimens and can be triggered by acute loss of PTEN in a p53-dependent fashion in mouse PCa models [33], suggesting that ERG activation might result in an initial senescent phenotype which requires subsequent environmental or genetic changes to progress to malignancy.

***ATG context and uORFs affect ERG translation efficiency and functions.***

Efficiently translated eukaryotic mRNAs require an optimal context around the start codon, to aid in its recognition by the ribosome (the Kozak sequence: GCCRCC**ATG**G) [34]. However, this is not always sufficient to explain translation efficiency. For example, although ERG-1a/b and T1:E4 share the same ATG in exon 4 (M4a) to initiate translation, their expression levels are different (Figure 7B lanes 2,3 and 5) suggesting a role for their different 5' UTRs. Indeed, substitution of the natural 5' UTRs with one containing a consensus Kozak (Figure 10A-B) led to activation of expression from the M3 ATG in the ERG-1b variant, with synthesis of the originally predicted 55 KDa peptide (Figure 10C, lanes 1-2). This confirms that elements within the 5' UTR of ERG-1b inhibit its translation and it rules out that the low ERG-1b (M3) abundance is due to an intrinsic instability of its N-terminal domain.

Upstream Open Reading Frames (uORFs) are typically short ORFs that start within the 5' UTR, are out-of frame with the main downstream coding sequence and can reduce its expression [35,36]. While no uORFs are present in ERG-1c or in T1:E4, three uORFs exist in the ERG-1b 5'UTR (Figure 10D, top), which could interfere with recognition of the predicted ATG of ERG-1b (Table 1). Abrogation of the uORF starting at uM2a by mutation of the ATG to GCG led to increased ERG-1b expression (Figure 10E, lane 2), indicating that the engagement of the scanning ribosome by the first encountered ATG, to generate a short 29 aa uM2a peptide, is a limiting factor in ERG-1b expression from the downstream ATG.

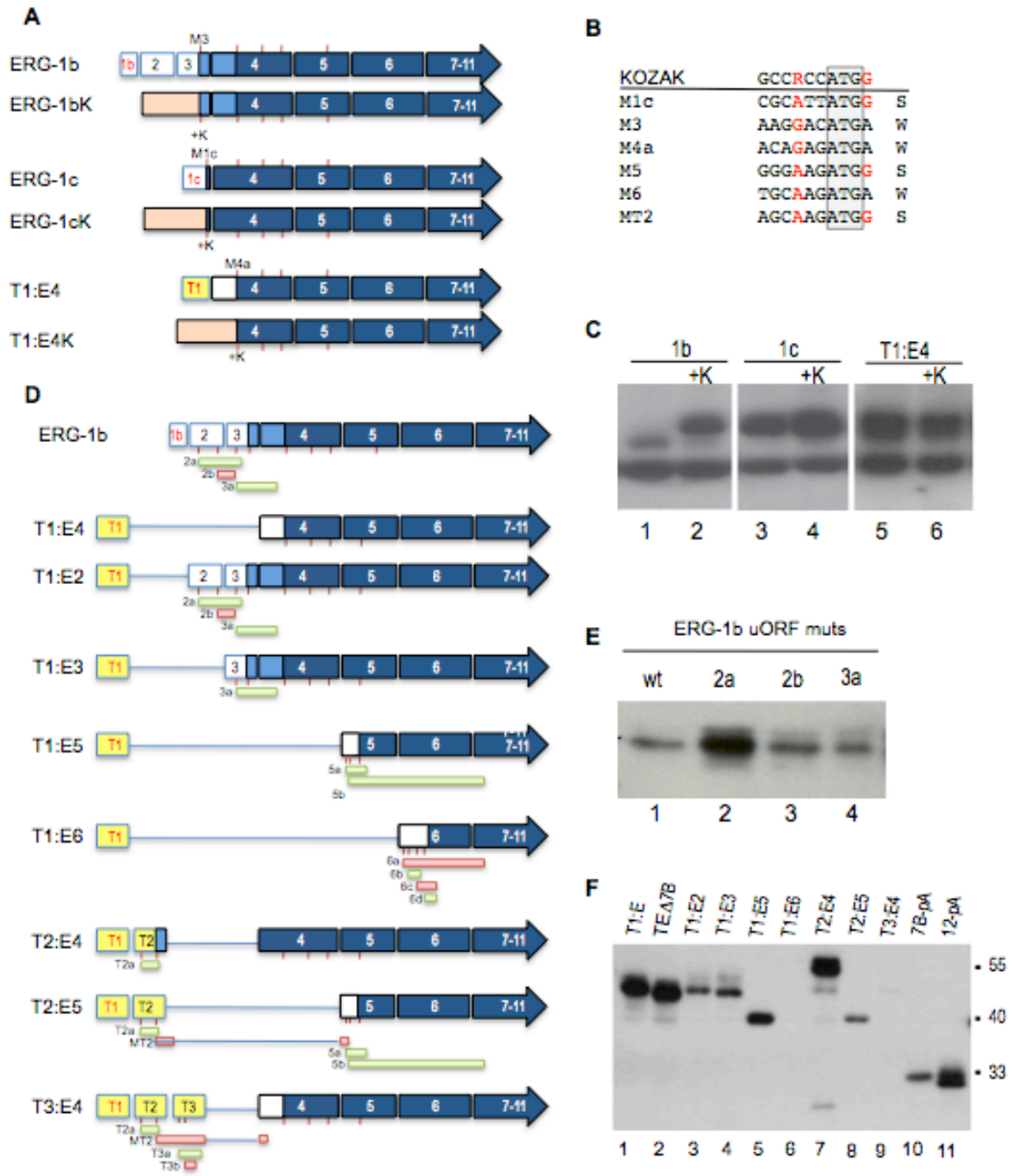



***Figure 10. Role of the 5' UTR region in ERG and Tmprss2:ERG variants expression (A)*** *The native 5' UTR of ERG-1b, ERG-1c and T1:E4 were replaced with a common one from the expression vector (in orange), and an optimized Kozak sequence. The replaced ATGs are in exon 3 (M3), 1c (M1c) and 4 (M4a)* ***(B)*** *Context of multiple in-frame ATG used as start codons in various ERG and Tmprss2:ERG variants. The sequence around the ATG is aligned to a consensus Kozak sequence, with the important conserved positions at -3 (R) and +4 (G) highlighted in red (counting the A as +1). Context is evaluated as 'strong' (S) if both positions are conserved, and 'weak' (W) if they are not. An expanded analysis of ATGs in the 5'UTR region of ERG and Tmprss2:ERG is reported in Table ST1.* ***(C)*** *WB analysis of the variants and mutants represented in (A). Improving the M3 context favors its use at the expenses of the M4a ATG.* ***(D)*** *uORFs that might influence translation efficiency of ERG variants in the 5' UTRs of ERG-1b and several common Tmprss2:ERG variants. Red ticks indicate ATGs, boxes below indicate the putative uORF generated and their approximate length (drawing not to scale). Red boxes indicate strong ATG contexts and green weak ones (as defined in (B)). More details about the uORFs are listed in Table ST1.* ***(E)*** *Effect of independently mutating uORFs' ATG in ERG-1b: mutation of the first ATG in exon 2 releases suppression of translation and increases levels of ERG from the ATG in exon 4.* ***(F)*** *Expression of ERG and Tmprss2:ERG variants. Full-length cDNAs for the variants, including their entire 5' UTRs, were expressed and lysates from transiently transfected HeLa cells were analyzed by western blot.*

The presence of uORFs could also similarly affect the levels of expression of the various androgen-driven Tmprss2:ERG fusion proteins, with prognostic implication, since the 5' heterogeneity in fusion is associated with different clinical outcomes [22,23]. The different pathological profiles could be due to changes in biological activity depending on the primary protein sequence or to differences in its abundance, *via* modulation of translation. Such variations might explain how, for example, the T2:E4 fusion is associated with a more aggressive phenotype [22,23].

To investigate whether translation efficiency plays a role in the activity of the fusion variants and to characterize their biological properties, we subcloned the 8 most common TMPRSS:ERG variants (T1:E2, T1:E3, T1:E4, T1:E5, T1:E6, T2:E4, T2:E5, T3:E4) [22,23], along with the exon 7b skipping variant in the common T1:E4 context (TEΔ7b), and the two C-terminal truncated variants (TE7bpA and TE12pA). All variants were transiently expressed in HeLa cells under identical conditions (Figure 10F).

Expression of the 54 KDa peptide from the T2:E4 isoform, which is associated with an aggressive phenotype, is robust (Figure 10F, lane7), as expected from a transcript that contains

| ATG | product | length (aa) | size (KDa) | distance from ATG (nt) | overlap | KOZAK RCC ATG G | strength |
|---|---|---|---|---|---|---|---|
| M1c | ERG1c | 479 | 54 | / | / | ATT ATG G | S |
| 2a | uORF | 87 | 3.3 | 89 | N | CAA ATG A | W |
| 2b | uORF | 39 | 1.5 | 69 | N | AAG ATG G | S |
| 3a | uORF | 84 | 2.9 | 19 | Y/N | TAG ATG G | W |
| M3 | ERG1a/b | 487 | 55 | / | / | GAC ATG A | W |
| M4a | various | 447 | 51 | / | / | GAG ATG A | W |
| 5a | uORF | 20 | 2.2 | 29 | Y | CTG ATG A | W |
| 5b | uORF | 123 | 15.2 | 25 | Y | TGA ATG C | W |
| M5 | Δ4, T1:E5 | 387 | 44 | / | / | AAG ATG G | S |
| 6a | uORF | 74 | 8.5 | 112 | Y | GCT ATG G | S |
| 6b | uORF | 10 | 1 | 68 | N | ACC ATG T | W |
| 6c | uORF | 22 | 2.5 | 62 | Y | AAT ATG G | S |
| 6d | uORF | 8 | 0.9 | 20 | N | TCG ATG G | W |
| M6 | T1:E6 | | | / | / | AAG ATG A | W |
| T2a | uORF | 10 | 1.2 | 45 | N | TCG ATG C | W |
| MT2 | T2:E4 | | | / | / | AAG ATG G | S |
| MT2 (T2:E5) | uORF | 9 | 0.9 | 28 | N | AAG ATG G | S |
| MT2 (T3:E4) | uORF | 87 | 9.3 | 55 | N | AAG ATG G | S |
| T3a | uORF | 43 | 4.8 | 144 | N | ACT ATG A | W |
| T3b | uORF | 40 | 4.4 | 144 | N | ACC ATG G | S |

***Context and characteristics of uORF and in-frame ATG in ERG variants.*** *ATGs from the 5' region of various ERG variants are listed. ATG in frame with ERG are indicated in red and preceded by a 'M' (=Met), ATG resulting in uORF are indicated in black, with the number indicating the exon that harbors them. The predicted length of the ORF/uORF (in amino acids and KDa) is reported, along with the distance (in nucleotides) from the translated ATG, whether the uORF overlaps with the translated ORF, and the ATG context. The context is considered 'strong', if both the determinant positions at -3 (G/A) and +4 (G) are conserved, and 'weak' if not.*

a strong native translation initiation site (MT2, Figure 10B). The presence of a weak overlapping uORF (uMT2a) is not sufficient to reduce T2:E4 expression. Isoforms T1:E2 and T1:E3 include the native ERG ATG on exon 3 (M3), and in principle should be efficiently expressed. However, as for the native ERG variants ERG-1a and ERG-1b, the M3 ATG is not used. Instead, usage of the M4 ATG in exon 4 leads to generation of the same ~50 KDa product described for ERG-1b. Similarly to ERG-1b, T1:E2 low expression levels might be due to the presence of the 3 uORF depicted in Figure 10D. Indeed T1:E3, which lacks exon 2 and the uORFs it carries, has a slight but reproducibly higher level of expression than T1:E2 (Figure 5F, lanes 4 vs. 3). Isoform T1:E5 uses the good Kozak in exon 5 and is expressed efficiently, despite the presence of two uORFs (5a and 5b). The same 44 KDa protein is also generated by T2:E5, but less efficiently, due to the presence of two additional uORFs (T2a and MT2). A similar scenario plays out for the T3:E4 variant tested, which uses the weak ATG in exon 4. In this case, the presence of 4 inhibitory uORFs almost completely abrogates translation of the ~50 KDa product. Skipping of exon 5 generates the rare isoform T1:E6, a putative protein lacking the PNT domain but containing the Ets and transactivation domain. However, its expression is suppressed due to the weakness of the Kozak associated to the in-frame ATG in exon 6, and the presence of 4 inhibitory uORFs.

From these studies we can conclude that overall, the levels of transient expression of Tmprss2:ERG isoforms appear to correlate with their 5' UTR context. The strength of the Kozak sequence and the number and relative strength of the uORFs both contribute to translation efficiency of the ERG and Tmprss2:ERG variants, and should be taken into consideration as an additional layer of complexity when assessing the oncogenic potential of the fusion variants observed in clinical samples.

### *TMPRSS2:ERG fusion variants biological activity*

Although the characterization of the ERG isoforms at the protein level enabled us to better understand the regulatory mechanisms underlying their expression, the primary source of ERG in prostate cancer is through the formation of the androgen-driven TMPRSS2:ERG fusions. Because some fusion isoforms are described to be more highly expressed, as well as some isoforms being associated with a more aggressive cancer phenotype, it was important to assess the various fusions isoforms isolated for any differences in biological functions.

We thus extended the previous experiments to a broader panel the various TMPRSS2:ERG fusion transcripts, to evaluate their biological activity and assess their functional significance in prostate carcinogenesis. All the main isoforms, indicated in Figure 11, were stably expressed and their invasion/migratory properties were assayed as above (Figure 11 A-D).

Furthermore, to directly measure the transcriptional activities, we subcloned the VE-cadherin and the matrix metalloproteinase-1 (MMP-1) promoters, which are two very well-characterized downstream targets of ERG (24-26) (Figure 11 E). Wound healing and a transcription dual-luciferase assays were used to assess the biological activity of the various fusion isoforms identified. The rates of wound healing by the specific fusion isoforms correlated fairly well with the luciferase activity generated. Although some isoforms have been described to be more tumorigenic than others, and some isoforms are more highly represented in distinct prostate cancers, it is difficult to speculate on the importance of the activity generated by one specific isoforms relative to another.

The fusion isoforms for which translation initiates in exon 4 (T1:E4, TEΔ7B, T1:E2 and T1:E3) or in exon 5 (T1:E5 and T2:E5) displayed similar rates of wound closure (Fig.11 C, lanes 2-6 and 9). Similarly, the luciferase activity generated was comparable, except in the context of the VE-cadherin promoter, which shows lower luciferase activity for T1:E5 and T2:E5 (Fig.11 E lanes 1-5 and 8). The moderate levels of luciferase activity of T3:E4, as well as 40% higher migration rate relative to the vector control, confirms that protein translation does occur and generates a functional peptide.

Similarly, T1:E6 for which protein detection was not possible, generated very low migration levels and luciferase activity, but still greater than the vector control. The two isoforms in which the Ets and TAD functional domains are lacking, TE:7BpA and TE:12pA, did not exhibit any significant rate of migration or luciferase activity.

The truncated fusion isoforms further displayed cell proliferation and cell viability measures similar to the vector control, relative to T1:E4 and TEΔ7B in which the functional domains are expressed .

Two additional isoforms, TE:7bpA and TE:12pA, which lack Ets and TAD functional domains, where included in the above analysis because they could potentially display modifying or even dominant-negative properties on ERG activity. Additionally, TE:7bpA RNA is overrepresented in

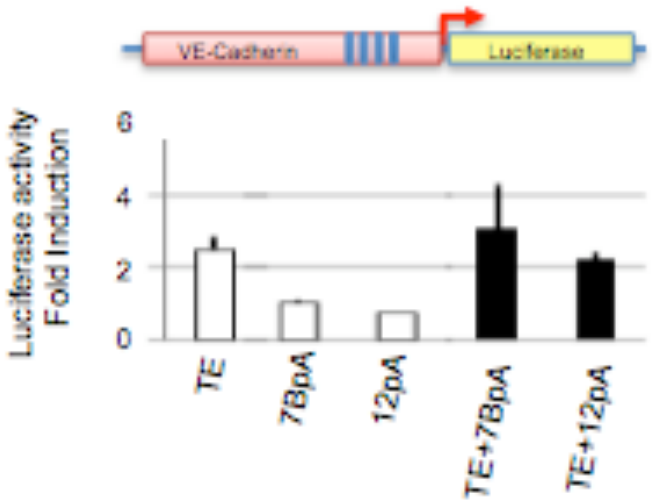



***Figure 12. Activity of truncated fusion isoforms lacking Ets and TAD domains.*** *Transient transcriptional activation of ERG-dependent VE Cadherin promoter. The luciferase reporter was transiently co-expressed in HeLa cells with full-length T1:E4 or/and truncated TE:7bpA and TE:12pA variants (plus a renilla luciferase vector to normalize for variation in transfection efficiency). Dual-luciferase assay was performed and activity is represented as fold-activity over that of co-transfected empty vector. Averages of at least 3 independent experiments, with standard deviations, are represented. The truncated isoforms alone cannot induce expression of luciferase, and when co-expressed together with the full-length protein they fail to inhibit its activity, ruling out for them a dominant-negative role. Further experiments would be required to reach more definitive conclusions.*

tumors. However, TE:7bpA is expressed at very low levels both in transient and stable experiments (Figure 10F, lane 10 and Figure 11A, lane 10), possibly because of intrinsic instability, making it impossible to reach any conclusion on its biological role. On the other hand, the robust expression of the structurally similar TE:12pA does not affect the rate of migration (Figure 6). In addition, it does not drive expression of luciferase from the VE-Cadherin or the MMP1 promoters, nor it can interfere with the activity of full-length variants in promoting such expression (Supplementary Figure 3). Thus, at least under these conditions, this isoform behaves like a null mutant rather than a dominant-negative, and the Ets and/or TAD domains encoded by exon 11 are essential for the activity of ERG variants.

### *Specific inhibition of oncogenic ERG variants*

We have shown that both the common PCa fusion variant T1:E4 and the native PCa-overexpressed variant ERG-1b use the same ATG in exon 4, whereas the variant most abundant in normal tissues uses one in exon 1c.

This raises the possibility to specifically target the cancer-associated ATG, without interference with the normally expressed ERG isoforms.

To this end, we decide to use phosphorodiamidate morpholino oligomers (morpholinos), a class of antisense compounds that can be used *in vitro* and *in vivo* to modulate gene expression by interfering with splicing patterns, miRNA maturation or translation [39,40]. Morpholinos targeted directly at the translation initiation site can inhibit translation by hindering recognition by the ribosome [39]. On the other hand, compounds targeted downstream of the ATG are easily displaced by the translating ribosome and are thus ineffective (Fig 13A).

In the case of ERG, a morpholino targeted to the exon 4 ATG should thus prevent its usage but not that of an upstream ATG, such as for example that on exon 1c.

In order to selectively block TMPRSS2:ERG and ERG-1b translation (since they share the same ATG), we designed two morpholino compounds: AS1, which targets the start codon in the fusion context (M4a), and AS2, which targets a sequence just upstream of the initiation ATG (Figure 13B).

Indeed, a compound (AS1) designed to bind to the region containing the ATG on exon 4 reduced expression of the T1:E4 variant in transient transfection experiments, whereas a similar compound (AS2) targeted to a region upstream did not (Figure 13B). Most importantly, neither had any effect on the transient expression of ERG-1c, the most commonly expressed variant in non-tumor tissues (Figure 13B).

These results suggest that development of specific translation-blocking compounds that effectively

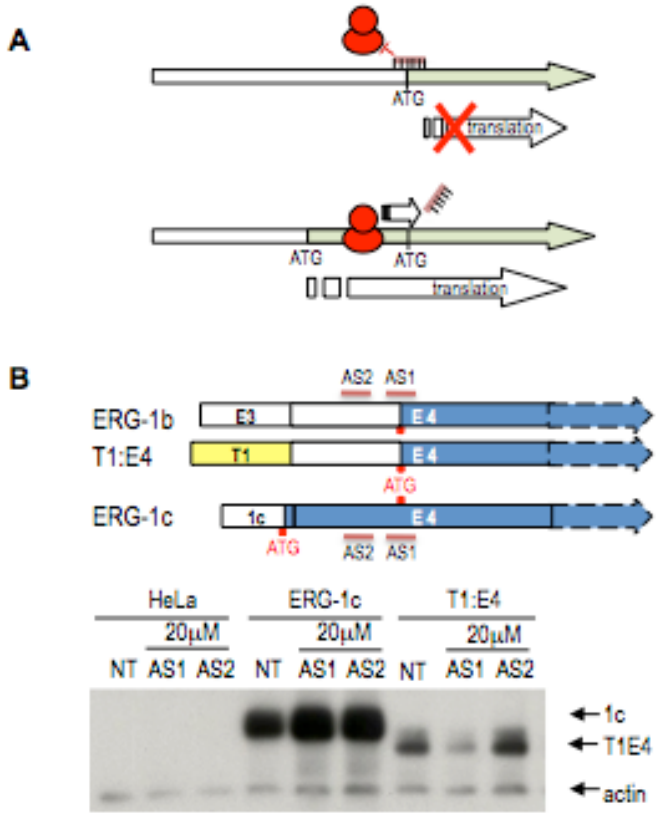



***Figure 13. Inhibition of Tempress:ERG initiation codon*** *(a) Predicted start codons in TMPRSS2 (exon 2) and ERG-1b (exon 3) are eliminated in the T1:E4 translocation. Mutational analysis of three potential in-frame ATG sites in T1:E4 identifies M4a as the fusion-specific start codon. (b) The morpholino strategy was used to inhibit ribosomal recognition of the ATG in the fusion-specific ERG without affecting the expression of normal ERG. Translation of T1:E4, and not ERG-1c, was specifically reduced by morpholino AS1*

and selectively reduce the levels of aberrantly expressed ERG isoforms without affecting normally expressed ones is feasible. Such compounds will serve as an important toolset to further our understanding of a pathway that is improperly activated in the majority of prostate cancers, and could form the basis of a future therapeutic approach.

We have devoted significant resources to test the AS1 compounds, in order to explore their anti-tumorigenic properties in vivo, by treating systemically and intra-tumorally xenograft models of ERG-overexpressing PCa, but we have unfortunately not been able yet to observe any efficacy of these compounds *in vivo*. Some additional experiments are in progress using funding from additional sources.

The failure to observe in vivo efficacy could be due to multiple reasons, which are currently under investigation. In particular it could be cause by ineffective delivery to the tumor and/or uptake by tumor cells; by efficient delivery to the tumor cell but ineffective efficacy in inhibiting ERG expression; or by the by efficient delivery and inhibition of ERG expression in tumor cells, but the lack of a significant biological effect on tumor development of ERG depletion.

The resolution of these question will shed important light on the significance of ERG expression in a PCa physiological context and on the viability of this approach as a future treatment.

## KEY RESEARCH ACCOMPLISHMENTS

1. **Clarification and new understanding of the genomic structure of the ERG gene**
2. **Characterization in normal tissues and prostate cancer of ERG variants expression**
3. **identification of endogenous ERG as a major source of Erg expression, in addition to Tmprss2:Erg**
4. **identification of a myc-driven regulatory forward loop that leads to endogenous ERG overexpression**
5. **identification of a truncated variant (7b-pA) as one of the most abundant isoforms in tumors**
6. **Mapping of the ATG starting sites in various variants, including the oncogenic ones**
7. **Role of the 5'UTR and of uORF in ERG isoform expression**
8. **Characterization of biological activity of various isoforms in vitro**
9. **ERG-dependent senescence**
10. **Specific inhibition of the oncogenic ATG in cells**

## REPORTABLE OUTCOMES

**Abstract/presentations**

George Boutsalis, Francesca Zammarchi, William Gerald, Luca Cartegni. **Oncogenic potential of multiple splicing isoforms of the prostate-cancer specific *TMPRSS2-ERG* gene fusion products.** RNA Society's 2010 Annual Meeting, Seattle, (2010).

Francesca Zammarchi, George Boutsalis, Luca Cartegni. **5' UTR control of native ERG and of Tmprss2:ERG variants activity in prostate cancer.** *PLoS ONE, under review* (2012) **.**

## CONCLUSIONS

Several attempts to correlate the expression of the TMPRSS2:ERG fusion to the clinical status of prostate cancer, or to characterize it as a prognostic marker have resulted in confounding outcomes. This suggests that the involvement of TMPRSS2:ERG in prostate cancer is complex and further studies would be required to elucidate its role and prognostic value.

During our investigations, we found that a subset of native ERG isoforms, in addition to Tmprss2:ERG fusion-derived ones, is strongly up-regulated in prostate cancer samples carrying the Tmprss2:ERG fusion, and that the expression of both native and fusion-derived ERG are influenced by the 5'UTR context, in particular by the presence of short inhibitory upstream ORFs (uORFs), which might have prognostic relevance. We also demonstrated that all most common cancer-associated ERG variants share use of the same ATG in exon 4, and that this can be specifically blocked by antisense compounds.

In addition, we completed a systematic analysis of ERG variants that resolves the confusing and often conflicting nomenclature and will result very useful to all investigators in the field.

The discovery of the Temprss2-ERG fusion in prostate cancer has changed the panorama of prostate tumor biology. Notwithstanding the rapidly increasing amount of information about the role and biology of this oncogenic fusion, much remain to be understood about its patterns of expression, functions and mechanism of action.

In particular little is still know about the specific roles of its many variants. Native ERG itself displays a complex pattern of expression, with multiple isoforms arising from the combinatorial usage of 3 main promoters, two main splicing events and three main polyadenylation sites. These combine to generate 30 main mRNA isoforms that encode for 15 different polypeptides, with variation both at the N-terminal and at the C-terminal, plus the alternative inclusion of 24 internal amino acids. This complexity is compounded by the fusion with tmprss2 (and other partners) which can occur at different points and is associated to complex alternative splicing patterns.

The work funded by this grant helps elucidating this complex scenario by rationally organizing and characterize ERG variants.

The observation that subset of native (non-fusion) ERG transcripts is strongly activated in tumors is also very important because it shifts the focus back to what contribution may come from the endogenous gene and underscores the possibility of a positive-feedback loop involving ERG and Tmprss2-ERG, which could conceivably lead to androgen-independence.

The identification of controlling 5'UTR elements, in particular the translation initiation context and the presence and characteristics of uORFs, might help better understanding the correlation between expressed variants and pathological characteristics. Besides, a correct understanding of what polypeptides are actually expressed from the different mRNAs would inform both in vitro and in vivo experiments with model animal systems.

Finally, we showed proof-of-principle evidence that the cancer-associated oncogenic ERG variants could be pharmacologically inhibited without disturbing the functions of the 'normal' variants expressed in non-tumor tissues. Although much more work needs to be done to confirm these results in vivo using PCa models (at which we were so far unsuccessful), It is a promising approach which could be in principle applied to other such oncogenic fusions.

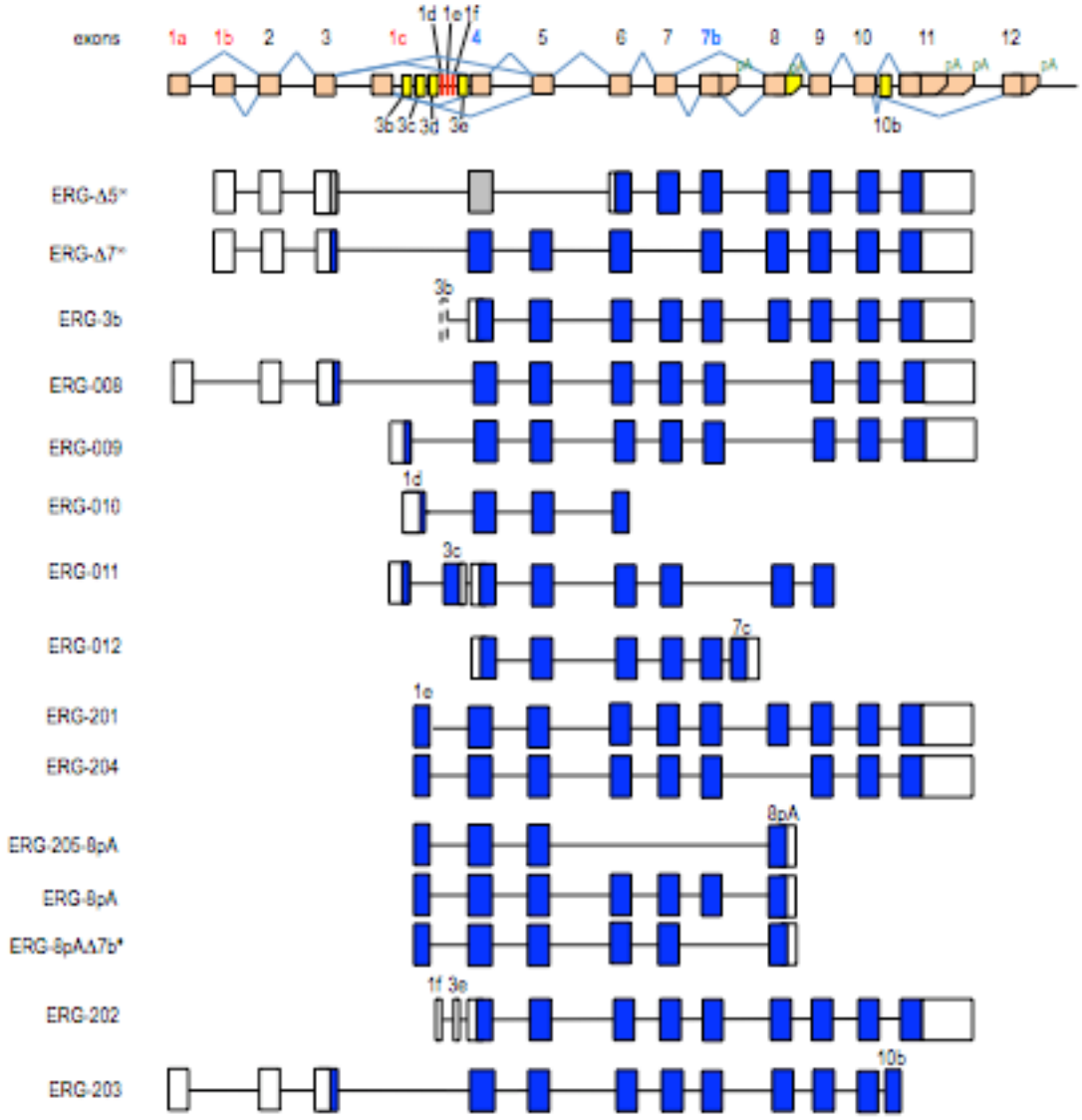
## REFERENCES


1. Rowley JD (2001) Chromosome translocations: dangerous liaisons revisited. Nat Rev Cancer 1: 245-250.
2. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310: 644-648.
3. Shand RL, Gelmann EP (2006) Molecular biology of prostate-cancer pathogenesis. Curr Opin Urol 16: 123-131.
4. Clark JP, Cooper CS (2009) ETS gene fusions in prostate cancer. Nat Rev Urol 6: 429-439.
5. Tomlins SA, Mehra R, Rhodes DR, Smith LR, Roulston D, et al. (2006) TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. Cancer Res 66: 3396-3400.
6. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, et al. (2008) Role of the TMPRSS2-ERG gene fusion in prostate cancer. Neoplasia 10: 177-188.
7. Helgeson BE, Tomlins SA, Shah N, Laxman B, Cao Q, et al. (2008) Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. Cancer Res 68: 73-80.
8. Carver BS, Tran J, Gopalan A, Chen Z, Shaikh S, et al. (2009) Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. Nat Genet 41: 619-624.
9. Demichelis F, Fall K, Perner S, Andren O, Schmidt F, et al. (2007) TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. Oncogene 26: 4596-4599.
10. Nam RK, Sugar L, Wang Z, Yang W, Kitching R, et al. (2007) Expression of TMPRSS2:ERG gene fusion in prostate cancer cells is an important prognostic factor for cancer progression. Cancer Biol Ther 6: 40-45.
11. Perner S, Demichelis F, Beroukhim R, Schmidt FH, Mosquera JM, et al. (2006) TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res 66: 8337-8341.
12. Rajput AB, Miller MA, De Luca A, Boyd N, Leung S, et al. (2007) Frequency of the TMPRSS2:ERG gene fusion is increased in moderate to poorly differentiated prostate cancers. J Clin Pathol 60: 1238-1243.
13. Klezovitch O, Risk M, Coleman I, Lucas JM, Null M, et al. (2008) A causal role for ERG in neoplastic transformation of prostate epithelium. Proc Natl Acad Sci U S A 105: 2105-2110.
14. Trotman LC, Niki M, Dotan ZA, Koutcher JA, Di Cristofano A, et al. (2003) Pten dose dictates cancer progression in the prostate. PLoS Biol 1: E59.
15. Squire JA (2009) TMPRSS2-ERG and PTEN loss in prostate cancer. Nat Genet 41: 509-510.
16. Rao VN, Papas TS, Reddy ES (1987) erg, a human ets-related gene on chromosome 21: alternative splicing, polyadenylation, and translation. Science 237: 635-639.
17. Prasad DD, Rao VN, Lee L, Reddy ES (1994) Differentially spliced erg-3 product functions as a transcriptional activator. Oncogene 9: 669-673.
18. Duterque-Coquillaud M, Niel C, Plaza S, Stehelin D (1993) New human erg isoforms generated by alternative splicing are transcriptional activators. Oncogene 8: 1865-1873.
19. Owczarek CM, Portbury KJ, Hardy MP, O'Leary DA, Kudoh J, et al. (2004) Detailed mapping of the ERG-ETS2 interval of human chromosome 21 and comparison with the region of conserved synteny on mouse chromosome 16. Gene 324: 65-77.
20. Carrere S, Verger A, Flourens A, Stehelin D, Duterque-Coquillaud M (1998) Erg proteins, transcription factors of the Ets family, form homo, heterodimers and ternary complexes via two distinct domains. Oncogene 16: 3261-3268.
21. Wang J, Cai Y, Yu W, Ren C, Spencer DM, et al. (2008) Pleiotropic biological activities of alternatively spliced TMPRSS2/ERG fusion gene transcripts. Cancer Res 68: 8516-8524.

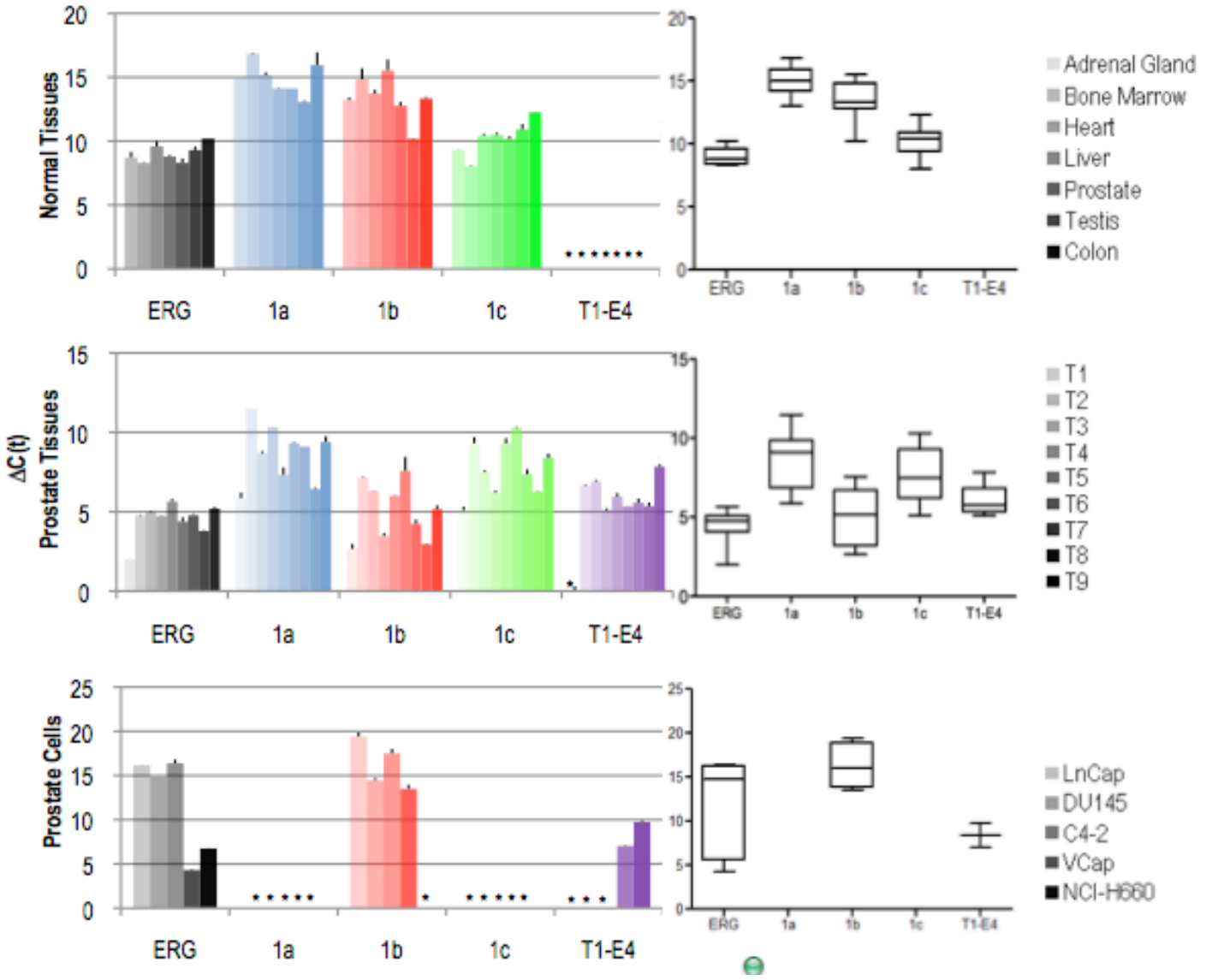
22. Wang J, Cai Y, Ren C, Ittmann M (2006) Expression of variant TMPRSS2/ERG fusion messenger RNAs is associated with aggressive prostate cancer. Cancer Res 66: 8347-8351.
23. Clark J, Merson S, Jhavar S, Flohr P, Edwards S, et al. (2007) Diversity of TMPRSS2-ERG fusion transcripts in the human prostate. Oncogene 26: 2667-2673.
24. FitzGerald LM, Agalliu I, Johnson K, Miller MA, Kwon EM, et al. (2008) Association of TMPRSS2-ERG gene fusion with clinical characteristics and outcomes: results from a population-based study of prostate cancer. BMC Cancer 8: 230.
25. Jhavar S, Reid A, Clark J, Kote-Jarai Z, Christmas T, et al. (2008) Detection of TMPRSS2-ERG translocations in human prostate cancer by expression profiling using GeneChip Human Exon 1.0 ST arrays. J Mol Diagn 10: 50-57.
26. Petrovics G, Liu A, Shaheduzzaman S, Furusato B, Sun C, et al. (2005) Frequent overexpression of ETS-related gene-1 (ERG1) in prostate cancer transcriptome. Oncogene 24: 3847-3852.
27. Mertz KD, Setlur SR, Dhanasekaran SM, Demichelis F, Perner S, et al. (2007) Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model. Neoplasia 9: 200-206.
28. Sun C, Dobi A, Mohamed A, Li H, Thangapazham RL, et al. (2008) TMPRSS2-ERG fusion, a common genomic alteration in prostate cancer activates C-MYC and abrogates prostate epithelial differentiation. Oncogene 27: 5348-5353.
29. Ghosh AK, Grigorieva I, Steele R, Hoover RG, Ray RB (1999) PTEN transcriptionally modulates c-myc gene expression in human breast carcinoma cells and is involved in cell growth regulation. Gene 235: 85-91.
30. Mani RS, Iyer MK, Cao Q, Brenner JC, Wang L, et al. (2011) TMPRSS2-ERG-mediated feed-forward regulation of wild-type ERG in human prostate cancers. Cancer Res 71: 5387-5392.
31. Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.
32. McDuff FK, Turner SD (2011) Jailbreak: oncogene-induced senescence and its evasion. Cell Signal 23: 6-13.
33. Chen Z, Trotman LC, Shaffer D, Lin HK, Dotan ZA, et al. (2005) Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. Nature 436: 725-730.
34. Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. J Biol Chem 266: 19867-19870.
35. Kochetov AV, Ahmad S, Ivanisenko V, Volkova OA, Kolchanov NA, et al. (2008) uORFs, reinitiation and alternative translation start sites in human mRNAs. FEBS Lett 582: 1293-1297.
36. Calvo SE, Pagliarini DJ, Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A 106: 7507-7512.
37. Buttice G, Duterque-Coquillaud M, Basuyaux JP, Carrere S, Kurkinen M, et al. (1996) Erg, an Ets-family member, differentially regulates human collagenase1 (MMP1) and stromelysin1 (MMP3) gene expression by physically interacting with the Fos/Jun complex. Oncogene 13: 2297-2306.
38. Birdsey GM, Dryden NH, Amsellem V, Gebhardt F, Sahnan K, et al. (2008) Transcription factor Erg regulates angiogenesis and endothelial apoptosis through VE-cadherin. Blood 111: 3498-3506.
39. Eisen JS, Smith JC (2008) Controlling morpholino experiments: don't stop making antisense. Development 135: 1735-1743.
40. Moulton JD, Jiang S (2009) Gene knockdowns in adult animals: PPMOs and vivo-morpholinos. Molecules 14: 1304-1323.

## APPENDICES

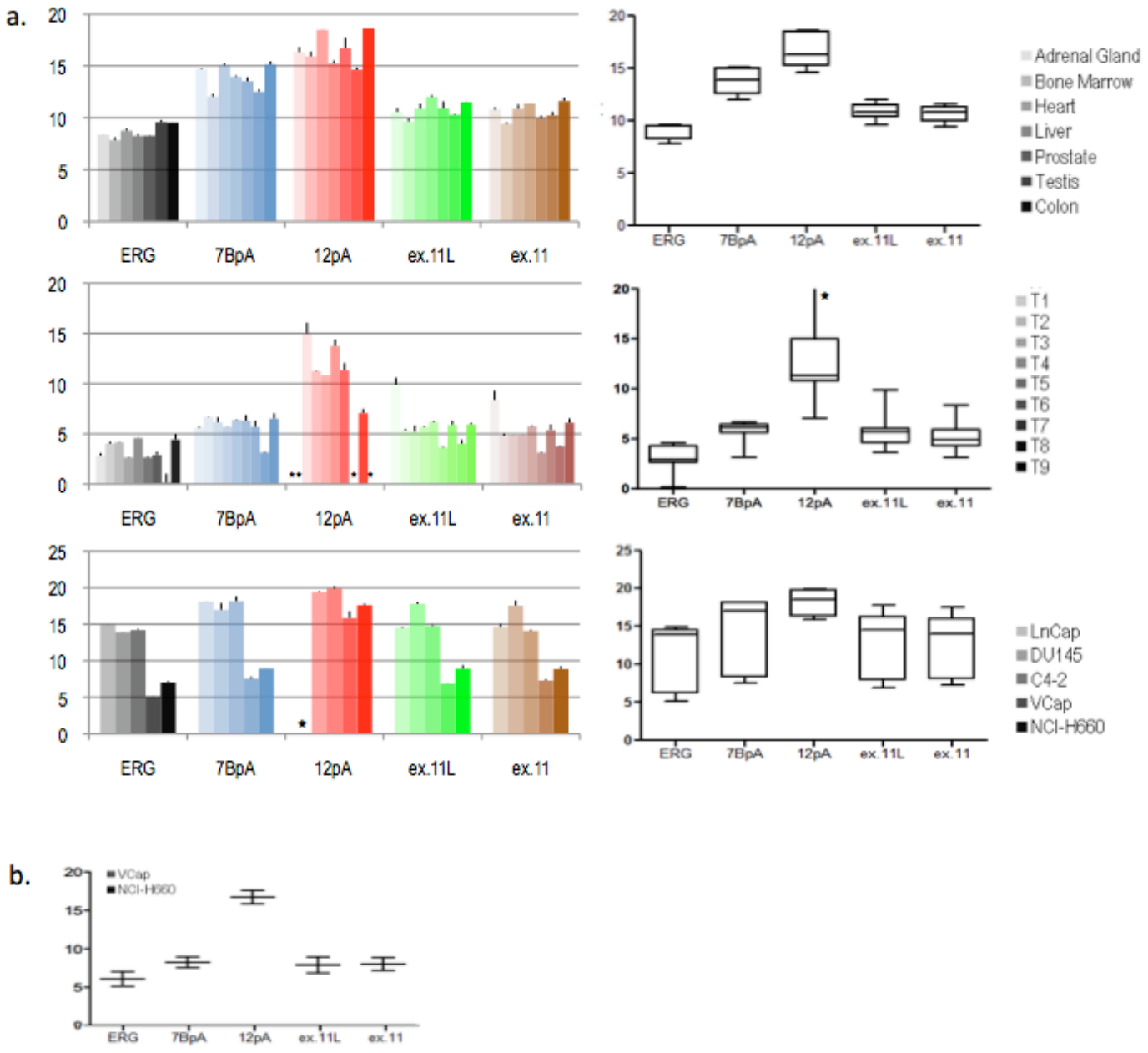
## SUPPORTING DATA



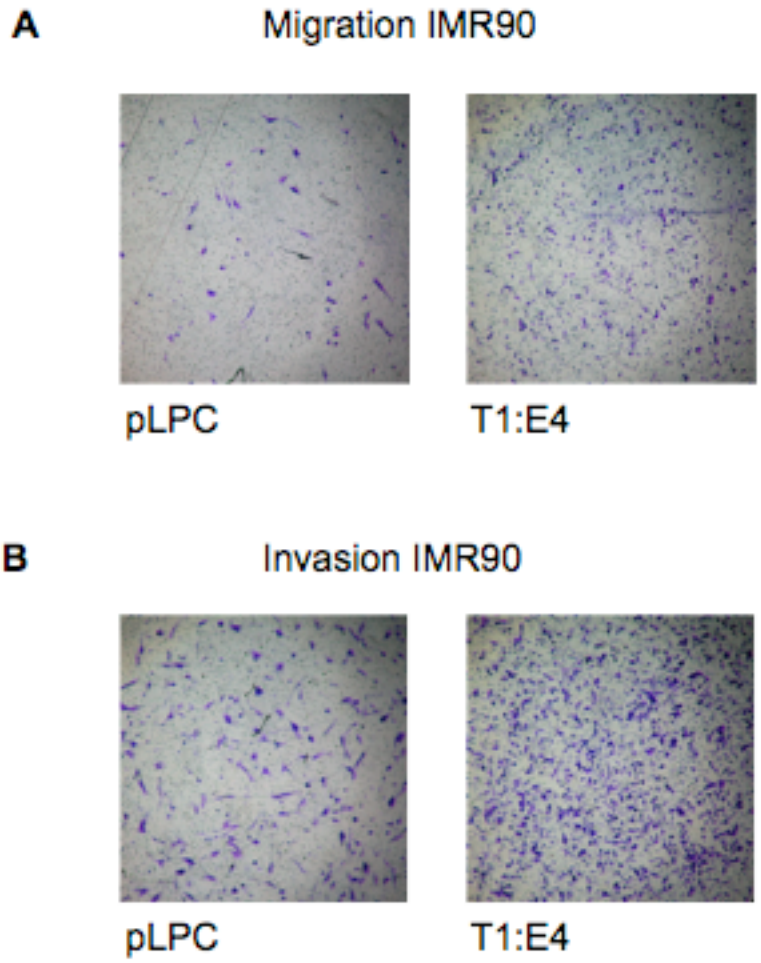

***Supplemental Figure 14. 'Minor' ERG Isoforms.***
*ERG exons and isoforms that have been described and listed in the Ensembl genome browser, or that we have identified (exons are shown in yellow and novel first exons are shown in red in the ERG locus schematic, top). The expression of these transcripts has yet to be proven (with the exception of the retention of intron 8 indicated in 8-pA). White boxes represent untranslated portions and those in blue represent the predicted translated proteins from the identification of in-frame ATG codons. Shaded region in ERG-Δ5 represents the introduction of a premature stop codon when exon 5 is skipped in the context of 1b.*
*(∞ The alternative splicing of exons 5 and 7 are shown in the context of 1b, but can also occur in the context of 1a and 1c. *Polyadenylation site 8-pA with the alternative splicing of exon 7B has been confirmed but not in the context of exon 1e.)*




***Supplemental Figure 2. Quantification of Alternative Promoter Use.***
*Promoter $P_C$ is the most active in normal tissues, while promoter $P_B$ is the most active in cancer tissue and the only one active in prostate cancer cell lines.*
*qPCR analysis of the alternative promoter use in a panel of normal tissues, prostate cancer samples and prostate cancer cells. Primer sets specific for the 3 different ERG promoters and for the TMPRSS2:ERG fusion where used, along with a primer set spanning exons 5-7 to quantify total ERG. The indicated values in the graphs on the left represent averages of 3 independent experiments and are presented as ΔC(t) normalized to GAPDH housekeeping gene, therefore a "high" ΔC(t) value means low levels of expression and a "low" value means high level of expression. Asterisks indicate that the gene was NOT detected in the sample and would therefore be equivalent to a column that goes to the top of the table, but it is omitted for clarity. The graphs on the right summarize the data on the left in a 'box and whiskers' representation with indicated the median and the 5th and 95th percentile value. NCI-H660 cells only express ERG from the TMPRSS2:ERG fusion because the fusion is present on both alleles and therefore the natural ERG promoters are completely absent.*




***Supplemental Figure 3. Quantification of Alternative Polyadenylation Sites usage***.
*The distal PolyA site on exon 11 (11LpA) is the most active in normal tissues, while PolyA site 7b is strongly activated in tumors and in prostate cancer cell lines that carry the fusion.*
*qPCR analysis of the alternative PolyA sites use in a panel of normal tissues, prostate cancer samples and prostate cancer cells. Primer sets specific for the 3 different polyA sites where used, along with a primer set spanning exons 5-7 to quantify total ERG and one set quantify total exon 11 levels to infer 11SpA usage. The indicated values in the graphs on the left represent averages of 3 independent experiments and are presented as ΔC(t) normalized to GAPDH housekeeping gene, therefore a "high" ΔC(t) value means low levels of expression and a "low" value means high level of expression. Asterisks indicate that the gene was NOT detected in the sample and would therefore be equivalent to a column that goes to the top of the table, but it is omitted for clarity. The graphs on the right summarize the data on the left in a 'box and whiskers' representation with indicated the median and the 5th and 95th percentile value.*




***Supplementary Figure 4. Effect of ERG 1b/T1:E4 ERG fusion variant on migration and invasion of IMR90 cells.*** *Following drug selection, ERG 1b/T1:E4 -overexpressing clones or empty-vector control (pLPC) were assayed for their migration and invasion potential using a transwell migration (A) or matrigel invasion assay (B).*